Peer-review report of

O'Mahony, C., Murphy, G., & Linehan, C. (2024). True discernment or blind scepticism? Comparing the effectiveness of four conspiracy belief interventions. *advances.in/psychology, 2*, e215691. https://doi.org/10.56296/aip00030

Round 1

Dear Authors,

Thank you for submitting your significant work to our special issue. I appreciate the evaluations provided by two experts in the field and extend my gratitude for their detailed and constructive feedback. In accordance with our journal's standards, I conducted an independent evaluation of the manuscript prior to reviewing their assessments.

We concur on the strengths of your work. The comparative testing of multiple interventions is commendable and tackles a prominent issue both in the field and socially. It is also noteworthy that the studies were pre-registered and appeared to be well-powered. However, we all have several major concerns that need to be addressed before considering your work for publication in the special issue. I will not reiterate the reviewers' points, as they have articulated their comments clearly and concretely. Nevertheless, I ask you to address each of their points in your revision and to outline how you did so in a revision letter, including excerpts that demonstrate the changes implemented. Additionally, please find below my own comments, which may overlap to varying degrees with those of the reviewers.

General Issues

Your research analyzes differences in two measures, one of which is the CTAS. This measure raises several questions and concerns. The first set of concerns is conceptual. One issue is that a correct score is assigned to a single answer, yet multiple answers could be correct simultaneously, based on the provided example (this issue also applies to the quiz in the active inoculation task). For instance, the statement, “This reasoning is flawed. Oil companies wouldn’t want to draw bad press on themselves,” seems like a reasonable heuristic (or “critical thinking”), though it might not directly address a logical flaw in reasoning. Many other items present similar challenges. In the “James Greenhouse” case (supplementary materials), with the limited information provided, it is not entirely clear whether he is a “singular academic.” Furthermore, it is ambiguous why “Dr. Greenhouse should be trusted; scientific breakthroughs often come from those questioning the status quo” is considered a “wrong” answer. This issue recurs in many other items. For example, in the UFO item, statements like, "The UFO sightings are likely just a distraction from

something more significant that the government doesn't want us to know. They're probably using this to divert our attention from more pressing issues," seem to tap more into political trust than “critical thinking about conspiracy assessment.” Clarifying what your measure assesses and what it does not would be beneficial (see similar points by the reviewers).

The second, and arguably more severe, issue with the scale is its unacceptably low reliability, which possibly ties back to the conceptual issues previously mentioned, though I recognize that a longer version produced better results previously. As it stands, in this study, the reliability is too low to permit any meaningful interpretation of effects. To address this, please conduct a factor analysis of the responses (recoded as 0 for false response and 1 for correct response) to identify which items cluster together or, conversely, exhibit weak loadings. Subsequently, use reliability analyses to exclude items with weak inter-item correlations. Unless acceptable reliability is reached, please omit any analysis of the scale. Additionally, you explore scores on “plausible” and “implausible” items. What is the validity (factor distinction) and reliability of these subscales? Without establishing acceptable reliability, I am afraid we cannot interpret these findings neither.

As it stands, the CTAC cannot be interpreted, and no statistically significant effects were observed on the reliable IEUB. However, this result is not adequately reflected in your paper. The general finding is that all conditions failed to produce meaningful effects on the measures that were reliable. Even if, based on factor analyses and new reliability tests, you can create better CTAC scales and find that effects hold, the IEUB findings constitute half of the results and should be given considerably more weight in the abstract and discussion (they indicate that the interventions were not very effective).

Substantial changes were also made to the IEUB. Please include a report of a factor analysis of the items. Additionally, this scale included five factual statements, which seem particularly well-suited for assessing discernment but were not analyzed. I recommend including these statements in your analyses, especially since the CTAS did not appear to perform effectively.

Like the reviewers, I expected an experiment in the second study, but this is actually a separate data collection that is then merged with the first (experimental) study. The second study does not qualify as an experiment due to the lack of random assignment and therefore cannot be compared to the other conditions from Study 1 in an unbiased manner. It should not be described as an experiment. As Reviewer 1 suggested, I recommend reporting it as part of Experiment 1 but being transparent of it not being part of the random assignment. You should thus report its sample separately and clearly indicate that this was not an experimental condition but

collected separately. Explain how the recruitment of participants differed, how you ensured that participants from the first experiment did not participate again, and conduct comparisons of the samples based on demographics. All comparisons to the other conditions need to be conducted in ANCOVAs that control for demographics. The results section focusing on the discernment condition should be significantly toned down in making causal claims. For instance, you cannot assert that the new condition “improved” something or had “effectiveness” due to the lack of random assignment; results are only suggestive.

As Reviewer 1, I was also concerned about the role of response bias. Some parts of the experimental manipulations (e.g., the quiz) seemed very similar to the CTAC. How do we know that the findings you observe are not influenced by response biases (e.g., experimenter bias)? This issue needs to be thoroughly discussed in your limitations section. Additionally, please make it clearer what specific fallacies the conditions and CTAS focused on (name all) and to what extent these overlapped.

Both the reviewers and I attempted to replicate your analyses, but the necessary code was missing. Please ensure that the analysis code is available in OSF. Additionally, make sure these files are anonymized. Currently, they seem to contain IP addresses, location data, and Prolific IDs. All of this identifiable information, as well as any other potentially identifiable information, needs to be deleted.

As Reviewer 2 pointed out, please ensure that you specify the type of test used in the power analyses (I assume it is an ANOVA).

Generally, please calculate average rather than sum scores to facilitate interpretation.

For Figure 3, please confirm that you have the copyright for this graph.

Please report findings with the pre-registered exclusion criterion in the main text and then the full sample in the SOM (rather than vice versa).

Please explain why different attention checks were used for different conditions. Please provide failed response rates per condition.

“A post-hoc Tukey’s test showed that, in an inverse of the results for the Implausible subscale, participants were significantly less accurate at critically appraising implausible conspiracy theories in the Active Inoculation condition than in the Priming ($p < .001$, $d = 0.48$) and Control conditions ($p < .001$, $d = 0.61$). Similarly, participants were significantly more accurate at critically appraising implausible conspiracy theories in the Inoculation condition than in the Priming ($p < .001$, $d = 0.35$) and Control conditions ($p < .001$, $d = 0.48$). There was no difference between Inoculation and Active Inoculation conditions ($p = .360$) nor the between Control and Priming conditions ($p = .445$).” These findings are confusing. You describe them as “similar”, when they indeed go in the opposite direction.

Stylistic issues

There are several linguistic and stylistic issues that need thorough attention. Please consider the following points only as examples. There are many more, and your manuscript requires a comprehensive proofreading.

- Please adhere closely to APA7 guidelines in terms of heading, figures & tables (titles and captions), statistics, and text formatting (each needs significant changes).

- In line with APA7, use "women" or "men" (or other gender) to describe the sample, rather than "male" or "female." If possible, provide more information about respondents, such as their education, political leaning, and ethnicity.

- Ensure that all numbers under 12 (unless referring to statistics) are written out in accordance with APA7.

- Italicize “*d*” as in Cohen’s *d*. Other statistical symbols should also be correctly presented (non-Greek letters should be italicized). Please adhere closely to APA7 guidelines in statistical reporting.

- Italicize scale endpoints in APA format, e.g., 1 = *I’m sure this is false*, to 5 = *I’m sure this is true*.

- You refer to Figure 2 for your results, but this figure presents something else.

- Example items or conditions that consist of text only should be presented in text, not as a figure. The same applies to the priming material, unless it consists/includes graphical components.

- Restructuring your methods section will be helpful. Present experimental manipulations as subheadings under procedures and present these procedures before measures. By this, the structure will be chronological.

- The first sentence of your manuscript is incomplete.

- The citation format needs revisions. In text, do not use "&" between dual authors (only use it within parentheses). I.e., use “and” instead (e.g., "Banas and Miller (2013)" but "as shown in previous work (Banas & Miller, 2013)").

- Page 9, line 11 contains a typo.

Please ensure that these and similar issues are thoroughly addressed in your revision.

To sum up, we see clear potential in your submission, but believe that it needs significant revisions, including some major reframing, restructuring, re-analyzing, to strengthen your statistical conclusions and enhance the impact of your paper. Thank you again for submitting your interesting work to our special issue. I look forward to receiving a revised version by the end of June. Please reach out to me should you need additional time.

Best regards,

Jonas R. Kunst

Editor-in-Chief

Reviewer 1:

The manuscript reports one (1) study testing and comparing different levels of an intervention aiming to increase discernment between plausible and implausible conspiracy theories. The levels of intervention range from a) a warning, b) warning + explanation of common biases, c) warning + explanation + quiz testing 5 implausible conspiracy theories, d) warning + explanation + quiz including 5 implausible + 1 plausible conspiracy theory. The study hereby was conducted in two stages, the first stage only including a) – c), the second stage being an additional recruitment to test d).

There are many things to like about the manuscript and the study. Notably, the study was preregistered, features open data and materials, and utilizes sequential testing across its studies. The interventions build upon each other, making a direct comparison possible. This kind of study design is what I would like to see more often in research. Additionally, the research question is both interesting and timely. Examining the belief in plausible and implausible conspiracy theories and developing trainings (interventions) to foster discernment is a significant gap in the literature that has not been adequately explored until now.

I wanted to replicate the analyses. The OSF files that were provided included the data, but not the syntax that was used to produce the results. Thus, I was not able to replicate the same analyses. I looked into the dataset nonetheless – and noticed that the N's for the conditions do not match the ones reported in the paper (i.e., the paper mentions N = 202 for the Discernment condition, while the data only includes N = 201). I reproduced the ANOVAs. Given the differences in the N, I noticed slight differences towards the paper, but the main conclusions were the same.

In addition, I have several remarks towards the paper:

1. Presentation of the study

The abstract and headlines give the impression that there are two distinct studies (potentially with the second one replicating the first). However, "Study 2" is actually just an additional data acquisition for the first study of N = 202. The data analysis uses the same participants as "study 1". Thus, it is not accurate to speak of these as separate studies. The authors may want to speak of it as two stages or additional data acquisition. I, personally, would also not be opposed to presenting it as one study overall (not even distinguishing the stages), mentioning the additional data acquisition and two preregistrations in the beginning, and only presenting analyses for the full dataset, reporting the analysis excluding condition d) in the supplement for transparency.

It is a limitation that overall results are only based on a single study, which could be mentioned as a limitation in the supplement.

There are also validity reports (reliability, M, SD) missing for the final scales. That is, while the reliability for the overall CTAC of stage 1 is reported, details are missing for the subdimensions (which are included in the analyses) as well as for reliabilities for stage 2.

2. Dependent Variable, Similaritiy to the "Active Inoculation" Intervention

The measure of the CTAT appears very similar, if not identical, to what has been used in active inoculation interventions, where participants choose between 1-4 options to evaluate biases. What is the difference between the two? What makes it legitimate to speak of one as intervention, while the other one tests the effectiveness of the intervention?

This raises questions about the effectiveness of the CTAT as a measure of discernment versus it being another repetition of the previous task. It is not surprising that participants in stage 2 scored better in this condition, as they may have been better trained for this type of test rather than genuinely improving their discernment. Ideally, another study should test the actual effect of discernment on a different measure, perhaps by confronting participants with both implausible and plausible conspiracy theories and then evaluating their beliefs. While additional data collection might not be feasible, at the very least, the similarity between the condition and the dependent variable should be better explained and discussed as a limitation.

Furthermore, results from measures other than the CTAT should be given more weight.

In this light, it is positive to note that the authors also include the Index of Epistemically Unwarranted Belief scale. Yet, the fact that the “interventions” do not produce any differences in this index makes me wonder whether there really is an effect on discernment abilities (rather than just testing effects).

I am also hesitant to interpret the results from the CTAT due to its low reliability. Did the authors conduct a factor analysis? See if there is one specific item that might be removed? A discussion of potential reasons for this low reliability should be included in the relevant paragraph of the paper.

3. Framing in Terms of Comparing Intervention Designs

The authors make a bold claim by stating that they are testing different intervention strategies that have been used in the literature. While this goal is commendable, the conditions of the study, especially those labeled “priming” and "inoculation," do not seem truly comparable to existing studies. For instance, the studies described as “priming” simply administer a warning, while “inoculation” includes the the explanation of biases. Studies typically using "inoculation" typically explain and warn people against a specific conspiracy theory they might be confronted with later. They typically provide participants with detailed background information (see e.g. Jolley, 2014). Of course, there are a variety of interventions out there by now, and the term “inoculation” is often used as an umbrella term for many different things. But this should be stated throughout, and it should be clear that authors are testing one of many ways of inoculation rather than testing the effectiveness of inoculation more generally.

4. Presenting the Effect Size of Priming

I am astonished that the papers of Swami et al. (2014) and Pennycook et al. (2020) are presented as “promising interventions”. Both have failed to replicate during the last years (Roozenbeek et al. 2021; Većkalov, B., Gligorić, V., & Petrović, M. B., 2024). Some might argue that Roozenbeek et al. 2021 did not really show a failed replication, but an effect size of d = .10, roughly equalling r = .05 is really not worth investigating. The authors seem to be aware of the failed replication by Roozenbeek (they cite him, though only in passing in the discussion). Yet, during the theory part, they cite both studies and mention that the overall effect sizes of interventions of this size are between d = 0.3 – 0.56.

This simply cannot be true taking into consideration all the studies mentioned by the authors.

Thus, the current presentation misleads readers by selectively highlighting effect sizes from small samples while only peripherally mentioning the replication study that included an extremely large sample size. Overall, I would not present Swami et al.

(2014) and Pennycook et al. (2020) and the like as promising interventions. I understand that researchers want to provide a good story, but this should not mean misleading other researchers or stakeholders which might only superficially read the paper, being left with the impression that indeed priming and accuracy nudges are promising interventions.

Given that previous priming studies have been ineffective, the findings of the authors are actually very much in line (not in contrast, as stated on p. 26, line 14/15) with previous findings.

5**. Minor remarks**

- the wording of the first sentence reads strange (“Conspiracy theories arguments put forward”)

- participant description: what effect size of 0.1 is it referring to? Is this cohen’s d? this should be specified.

- p. 16: The whole bottom paragraph is mixed up. The authors start with “the plausible CTAC subscale”, but when talking about the findings they talk about implausible conspiracy theories. I think this is just a mix-up in wording (i.e. the authors meaning plausible conspiracy theories rather than implausible ones), but this should be checked again.

- similarly p. 21, line 11: the authors first talk about the overall CTAC score, but then only the implausible conspiracy theories (while the CTAC includes both, plausible and implausible CTs, no?)

Reviewer 2:

I enjoyed reading this paper comparing four different interventions aimed at improving critical thinking about conspiracy theories. The authors compare the effectiveness of three different types of existing interventions (priming, passive inoculation, active inoculation), as well as a newly developed discernment intervention, in improving critical thinking about plausible and implausible conspiracy theories and reducing implausible beliefs. The authors claim that the three extant interventions increase the rejection of implausible conspiracy theories, but also do so for plausible conspiracy theories (i.e., they induce undiscerning scepticism). The discernment intervention is argued to circumvent this issue, performing similar to the inoculation interventions regarding critical thinking about implausible conspiracy theories and similar to the control condition regarding critical thinking about plausible conspiracy theories. No evidence for effects on implausible beliefs was found. Based on these findings, the authors argue that discernment between plausible and implausible conspiracy theories should be assessed more widely in research on the effectiveness of interventions aimed at reducing susceptibility to conspiracy beliefs.

The presented research has high practical relevance with regards to tackling susceptibility to belief in conspiracy theories. I believe the paper contributes to the literature on interventions tackling this susceptibility in several ways. Firstly, it provides critical tests of the effectiveness of existing interventions. Secondly, these existing interventions have received much attention in research on misinformation, but less in the context of belief in conspiracy theories, hence the application of these interventions in the context of conspiracy beliefs provides a valuable addition. Another laudable contribution is the focus on discernment (in both the utilised measures and the newly developed intervention), which has so far been neglected in research on conspiracy beliefs. Overall, the authors provide a valuable perspective on interventions tackling conspiracy beliefs informed by the related field of misinformation research. However, I have several concerns with the current version of the manuscript, which is why I advise a revision. There are some major concerns, which I would like to see addressed before suggesting the paper to be published, and some minor issues and comments. I believe addressing the minor concerns is not crucial but doing so could improve the quality of the manuscript.

Below is a summary of the major concerns I have:

Theoretical background:

1. While this point refers to an aspect of the methods employed by the authors (the critical thinking about conspiracies [CTAC] measure), I believe that it requires attention in the theorising of the manuscript. The CTAC is introduced as an alternative to measures of general tendencies to believe in conspiracy theories and measures of belief in specific conspiracy theories as outcome measures in studies evaluating the effectiveness of interventions. However, it is technically not a measures of belief in conspiracy theories, but critical thinking about them, which are two different constructs. For instance, individuals could indicate that the reasoning behind a conspiratorial narrative in a CTAC scenario is flawed, but still form beliefs in the underlying conspiracy theory. In its current form, the manuscript does not explain the relationship between critical thinking about, and belief in, conspiracy theories. Given that it is analysed as an outcome measure parallel to the measure of implausible beliefs and given that the authors refer to prior research examining intervention effects on belief measures to inform their predictions about the CTAC, readers may assume that CTAC is conceptualised as an outcome that is parallel to beliefs. However, it could also be argued that critical thinking about conspiracy theories is an antecedent, or even a moderator, of belief in conspiracy theories. I am not suggesting that the authors should conceptualise CTAC as an antecedent or moderator of beliefs, but I believe that clarifying the role of the CTAC in relation to beliefs would add clarity to the authors' theorising. Furthermore, the authors argue that the CTAC is potentially superior to measures of belief in specific conspiracy theories or measures of general tendencies to believe in conspiracy theories that have been used previously in the literature because it allows for the assessment of discernment. While this certainly is an advantage of this measure, the same end could also be achieved by measuring belief in specific

true/plausible and false/implausible conspiracy theories rather than critical thinking about conspiracy theories. Since the authors argue (rightly so) that the use of different measures of conspiracy theories being used in the literature makes it difficult to compare the effectiveness of interventions (page 4, lines 4-12), the introduction of yet another measure should be justified. To summarise, I believe the manuscript would benefit from highlighting the merits of assessing critical thinking about conspiracy theories beyond allowing for the assessment of discernment (which could also be achieved using belief measures), which also entails a clarification of the role of the construct in relation to beliefs so that readers can assess the conceptual value of using the CTAC as an outcome measure in testing the effectiveness of interventions.

2. Throughout the article, the authors refer to concepts, findings, and literature from both research on misinformation and research on belief in conspiracy theories, which I see as a valuable contribution, as I believe that these two related literatures can complement each other. And while some findings may translate from one domain to the other, this is not necessarily a given for all findings. Providing a theoretical argument for why the findings from the misinformation literature the authors draw on to inform their predictions may translate to the domain of conspiracy beliefs would strengthen the theoretical background of the paper. In the absence of such theoretical arguments, the authors could also highlight that examining whether concepts from misinformation research translate to a conspiracy context is in itself one of the aims (and hence contributions) of the paper.

Methods:

1. “Experiment” 2 is technically not a separate experiment, but an additional condition from a separate data collection that is compared to the conditions of Study 1. While this is acknowledged by the authors, the issues with this approach are not acknowledged. For instance, if the data collection for this additional condition was indeed separate from the other conditions as reported by the authors, participants could not have been randomly assigned to the additional condition and therefore participants in this condition may differ on various (unobserved) variables introduced by this design choice. The fact that the discernment condition outperformed the other conditions could therefore either be caused by the intervention being more effective or by participants being more discerning/engaging in higher levels of critical thinking than those in the other conditions. While there would be no reasons to entertain this possibility in a properly randomised design, this cannot be ruled out in this case, as the sample for the additional discernment condition may be from a different population than the sample from the other conditions. As no details about this additional data collection are reported, readers can also not judge any similarities or differences between the samples. For instance, no information is provided about when and how responses were collected for this additional condition - based on the data provided by the authors, it appears as if responses for the discernment condition were collected at the

same time as some responses for the other conditions. I urge the authors to elaborate on these details more thoroughly. It is also not apparent from the abstract that the sample of “Study 2” consists to a large part of the sample from “Study 1”. To avoid any confusion, this should be made clear throughout the text. However, most importantly, the potential limitations of this approach, including any potential differences between the two samples, should be acknowledged in the discussion of limitations, which is not the case in the manuscript in its current form.

2. For “Experiment 1”, the reporting of the power analysis states that “1095 participants were required to detect an effect size of 0.1 with 80% power”. I recommend that the authors report how the power analysis was conducted (e.g., which program or package), as well as which effect (e.g., omnibus effect, interaction effect, contrast between specific cells), and which effect size measure (Cohen’s *d*, *f*) the reporting refers to.

3. The measurement procedure of the CTAC described in the methods section of “Experiment 1” is a bit unclear to me. It is stated that there are four possible responses for each conspiracy scenario and that scores range from 1 to 8 (page 7, lines 1-3). I might have missed something, but it is not clear to me whether the rating from 1 to 8 applies to all four response options or only the one that is selected. It would be good to explain this procedure in more detail, mention what the scale endpoints (i.e., 1 and 8) refer to and to add the full reference to the CTAC paper the authors refer to (O’Mahony et al., 2024) to the bibliography, as it is currently only cited in text and not in the bibliography.

4. The authors make claims about discernment in individuals’ responses to conspiracy theories throughout the manuscript, but the analyses they report only allow for limited conclusions about discernment. For instance, the authors state that the negative effects of the non-discernment interventions on reasoning about plausible conspiracy theories were equal to the positive effects they had on reasoning about implausible conspiracy theories (page 26, line 24 and page 27, lines 1-3). However, the analyses reported by the authors do not allow for inferences about the significance of differences between responses to plausible and implausible conspiracy theories (i.e., discernment). More specifically, the authors conduct ANOVAs and multiple comparisons on both responses to plausible and responses to implausible conspiracy theories separately, which does not allow for statements about the significance of differences between these two types of statements. The tests conducted by the authors do not indicate whether the differences between responses to plausible and implausible conspiracy theories are significantly different from 0, as mere differences in significance or signs of effects for plausible and implausible conspiracy theories obtained from separate tests do not indicate significance of differences between these effects. Since one of the main contributions of the manuscript hinges on discernment, I recommend that the authors rectify this by using an analytical approach that quantifies differences between effects for plausible and those for implausible conspiracy theories (i.e., discernment) either in addition to, or instead of, the analyses

currently reported in the manuscript. For instance, this could be done by fitting a linear mixed model that allows to test whether the plausibility of conspiracy theories (as a repeated-measures moderator) moderates the effects of interventions. If I am not mistaken, the difference between intervention effects for plausible and implausible conspiracy theories could also be tested by fitting a multivariate model with responses to plausible and implausible conspiracy theories as separate outcome variables and conducting a linear hypothesis test to test whether the coefficients differ between the two outcome variables (e.g., Fox et al., 2013; *The R Journal*). The authors could also utilise analytical approaches that are commonly used in the misinformation literature to measure discernment between true and false information, such as computing discernment scores as an outcome variable or using methods from signal detection theory (see e.g., Higham et al., 2024; *Current Opinion in Psychology*).

Clarity:

1. The reporting of results was somewhat unclear in parts. For instance, in reporting analyses for the plausible and implausible subscales of the CTAC for “Experiment 1”, I believe the authors confused “plausible” and “implausible” (e.g., page 16, lines 20 and 22), making it difficult to follow. This should be corrected. Also, on page 21, starting in line 5, the authors report that an ANOVA was conducted on the overall CTAC scores. Then the authors report multiple comparisons, without specifying which scores (overall, plausible, or implausible, CTAC scores) these comparisons were run on. Lines 10 and 11 state “participants were significantly more accurate at critically appraising implausible conspiracy theories […]”, suggesting that comparisons were run on implausible CTAC scores. If that is the case, it should be clarified why different scores were used for the ANOVA and multiple comparisons and if not, this should be corrected to ensure interpretability of the reported results.
2. While the manuscript is fully understandable, it contains some spelling, punctuation, and formatting errors. I recommend that the authors check the manuscript for these errors. I listed some errors I spotted here (note that this is not an exhaustive list):
    1. Unless I am misunderstanding something, there seems to be a spelling error in title: it reads “discernment ‘of’ blind scepticism” rather than ‘or’. Also, I think either the question mark or the semicolon in the title can be omitted.
    2. Page 1, Line 1: An “are” is missing; “conspiracy theories *are* arguments”.
    3. Page 3, line 8: I believe the “and” is not needed here.
    4. It is unclear what “b, c, d” refers to in the in-text citation of Swami et al. (2014) on page 2 (line 8).

5. There is a spelling mistake in the example item on page 7 (“effected”).
6. Page 2, line 2: The subheading is in the same line as the text from the previous paragraph.
7. Page 7, lines 4 and 5: There are some punctuation errors.
8. Instead of inserting the example items as figures, I would simply insert them as plain text.
9. The title of Figure 3 is present twice in the text: once on page 12, referring to it as Figure 3 and once on page 13, referring to it as Figure 1.
10. Figure 6 is not referred to in the text.

Open science & reproducibility:

1. While the authors made the data and survey materials openly available, which is very positive, no analysis code or details needed to reproduce the reported analyses (e.g., which software, packages, functions were used) is provided (or I did not find it on the OSF page). I encourage the authors to provide the code that was used to conduct any analyses that are reported or details that would allow other researchers to reproduce them.

Hereafter, I summarise some minor issues and comments:

Theoretical background:

1. In introducing priming interventions, the authors refer to the studies conducted by Swami et al. (2014) showing that difficult-to-read fonts reduce susceptibility to conspiracy beliefs. The authors then also mention that this approach has failed to replicate by referring to Janouskova et al. (2022), which tested a similar method in a different context than conspiracy beliefs. I want to direct the authors’ attention to a direct replication of the Swami et al. (2014) studies in question in the form of a registered report by Većkalov et al. (2024; *Journal of Experimental Social Psychology*) that provides additional (even stronger) evidence for the point the authors make.
2. I am intrigued by the findings that some of the presented interventions appear to induce undiscerning scepticism and think this finding would warrant a theoretical discussion of the psychological processes by which these interventions may induce undiscerning scepticism. While I am not aware of any research explicitly testing such mechanisms, there may be merit in discussing this finding pertaining to undiscerning scepticism through the lens of research on factors leading to the undiscerning acceptance of conspiracy theories. For instance, the need for quick and definitive judgments (need for cognitive closure; e.g., see Marchlewska et al., 2018, *European Journal of Social Psychology*) and the need to judge whether things are entirely good or bad (need to evaluate; e.g., see Mang et al., 2024, *Personality and Individual Differences*) are not only both related to conspiracy beliefs and their consequences, but also both related to a tendency to make undiscerning

judgments (e.g., uncritical across-the-board acceptance/positive evaluations or rejection/negative evaluations of attitude objects of a category). By only presenting negative information about conspiracy theories (i.e., “all conspiracy theories are false”), the aforementioned interventions may lead to the undiscerning rejection of all conspiracy theories (regardless of their plausibility) by tapping into the same psychological tendencies that usually result in the undiscerning acceptance of conspiracy theories when individuals are exposed to them. While this is just one of many potential (speculative) explanations of these findings, I think a discussion of the processes that may lead to undiscerning scepticism would be highly interesting.

3. The authors define “unreasonable conspiracy beliefs” as “believing in a conspiracy that is very unlikely to be true, for example due to a lack of evidence, or due to the complexity of explanation put forward” (page 1, lines 6-8). An explanation of why the complexity of an explanation makes a conspiracy theory unreasonable (e.g., drawing on research on the conjunction fallacy) would be informative. Also, although this is a very minor technicality, I believe “improbable”, rather than “unreasonable”, would be a better description of these conspiracy theories.

Open science & transparency

1. First of all, it is laudable that the authors preregistered the presented studies and analyses and acknowledged deviations from the preregistered plans. However, there are a few minor unacknowledged deviations:
    1. For “Experiment 1”, the authors present a hypothesis about the effects of active participation in dichotomous terms (page 5, lines 21-23: “participants who receive interventions that require higher levels of active participation (active inoculation) will report higher critical appraisal of conspiracy theories compared to those who receive passive interventions (priming, inoculation)”. The corresponding hypothesis in the preregistration, however, frames this prediction in more continuous terms (“As active involvement increases across four conditions (Control, Nudge, Inoculation , Active Inoculation) critical thinking of conspiracy theories will increase”). The prediction in the preregistration could be interpreted as a prediction whereby critical thinking scores will follow this pattern: active inoculation > inoculation > nudge > control, rather than the pattern indicated by the framing of the hypothesis in the manuscript: active inoculation > priming & inoculation. This is just a minor technicality about one hypothesis being worded in more specific terms than the other I wanted to bring to the authors’ attention.
    2. For “Experiment 2”, the sample size reported in the manuscript differs slightly from the preregistered sample size, which should be acknowledged and, if there are reasons for this deviation, justified.

Overall, despite the concerns I have raised, I enjoyed reading the manuscript and believe that it has the potential to be theoretically and practically impactful when it comes to tackling conspiracy beliefs. I hope the authors find my comments helpful in further highlighting their contributions and strengthening the evidence and arguments put forward in the manuscript. I wish the authors all the best for their future work on this manuscript.

Round 2

Dear Authors,

Thank you for submitting your revised manuscript. The Reviewers and I appreciated the changes you implemented, particularly the addition of a new, fully randomized study, which significantly enhances the credibility of your results as well as additional analyses (e.g., calculated d scores). However, we were surprised to find that several of our previous comments were either ignored or only superficially addressed. Therefore, I request that you engage more thoroughly with these and the new requested changes in your next revision. Please be aware that the expectations for this revision will be higher, given the insufficient response to many of our points in the previous version. The next submission will be evaluated strictly on whether it comprehensively and in-depth addresses all our comments. This evaluation will inform my decision on whether to continue considering your manuscript for publication in the special issue.

As in the previous round, the Reviewers provided very detailed and constructive input, and I thank them for their time and efforts. I independently read the manuscript, so my comments overlap to some extent with their evaluations. Please make sure to thoroughly address every point.

ABSTRACT

Please add the average effect sizes or ranges of the effects you observed to the abstract, as this is information that most readers will find interesting.

INTRODUCTION

The presentation of the accuracy priming research still needs more balance, as both reviewers highlighted. The evidential value of studies should be weighted by their sample size and overall quality.

Given the somewhat unusual structure of the paper (e.g., the addition of Study 1b), it would be helpful if you provided a road map and explanation in a "The Present Research" section at the end of the general introduction. This will prepare readers for the remainder of the article.

METHODS AND RESULTS

I agree with the Reviewer's observation that the IEUB statements on factual information are suitable benchmarks for evaluating blind skepticism. Please proceed with conducting the requested analyses. This is especially critical as the CTAC had inacceptable reliability in the first study (see next point).

As stated before but for some reason ignored, a scale with a reliability of .54 cannot be meaningfully analyzed. As stated in the first review, the CTAC should therefore be omitted from the first study, noting the reason for this in the main text. You did not provide any factorial evidence for it, despite this being requested, which is surprising. Additionally, you did not provide alphas for the two subscales that are later analyzed. Unless these alphas are acceptable, the analyses are of little value.

Although you were asked to report the reliability estimates for your scales for the separate dataset (now Study 1b), you have not done so. Please address this.

Thank you for providing attention check numbers. However, the means of failed responses are not very informative. Please provide percentages or Ns instead. If these are unequal, this should be discussed as a limitation (see Reviewers' comments).

On page 23, you mention testing for the influence of demographic information but do not provide details on how this was done. Additionally, you report ANOVAs instead of ANCOVAs. It's essential to control also for differences in country origins and education levels (if available), beyond age and gender. Please include these controls directly in your analyses.

In Study 1b, it is sufficient to report only the percentage/N of failed attention checks for the new condition rather than to repeat the values for the conditions from Study 1a.

In response to my comment in the previous review, you explained that mean scores were created, but that seems still not to be the case. For example, on page 16, the

IEUB scale should have a 1 to 5 mean score, but it has a 0 to 100 score, indicating that you summated items. As previously stated, please use mean (average) scores for all your scales as they are easier to interpret. This applies to other scales and other studies as well (e.g., the GCUBS in Study 2).

In your figures, please indicate significant differences with comparison lines to guide understanding and include the p-values on these lines. Ensure that the figures are presented in lossless SVG format. Increase the size of the axis labels for better readability. Additionally, it is unclear why the CTAC implausible scale ranges from -2 to 10 in one of the figures, as -2 seems outside the range; please adjust this scale accordingly.

Please make sure to address the data inconsistencies observed by Reviewer 1. Additionally, respond to Reviewer 2's request by meticulously justifying and reporting any exclusion of outliers.

DISCUSSION

“To our knowledge no previous intervention study has utilised an outcome measure that allows the assessment of discernment (O’Mahony et al., 2023).” This statement is too broad, as there are several misinformation studies that test this, as you earlier reported in the paper. You probably mean conspiracy theory studies, but given that many “misinformation” items measure conspiracy beliefs, there are likely several studies that would go against the statement.

“In fact, the negative effects these established interventions had on participants’ ability to reason about plausible conspiracy theories was equal to the positive effect they had on participants’ ability to reason about implausible conspiracy theories. We propose that rather than improving participants’ ability to appraise new information and come to a correct conclusion, these interventions in fact encourage blind scepticism.” This is a very general statement that does not adequately reflect the nuances between conditions that you observed. I don't think it is very helpful. What equality are you referring to and was it tested? I would prefer to see the discussion focus more on the differences between the various interventions. Moreover, to facilitate a higher-level interpretation, you could more frequently summarize how many of your conditions (e.g., x out of x) were different from the control for a given test, as this is the most important comparison.

“Priming interventions were found to be no more effective than controls at improving critical appraisal of conspiracy theories or reducing conspiracy and unfounded

beliefs. This is consistent with the mixed literature regarding the effectiveness of priming interventions (Epstein et al., 2021; 38 1 Fazio, 2020; Roozenbeek et al., 2021; Swami et al., 2014; Većkalov et al., 2024).” This is unclear. How can something be consistent with a “mixed literature”? I think you mean consistent with the part of the literature that fails to observe effects. This point relates to the description of this field in the introduction.

STYLISTIC ISSUES

This again is repetitive as many of these statements were made already in the previous round. I was surprised by not seeing them addressed thoroughly.

There are still typos, even in the first paragraph, such as a missing "as" in "such as." Please use a tool like Grammarly to identify and correct all typos, missing words, etc.

As previously stated, ensure that ALL non-Greek statistical letters are italicized.

Please ensure that p-values are always reported to three decimal places, using "$p <$ .001" if the value is under .001. Additionally, report all other statistics to two decimal places (this varies somewhat in your text).

Although you confirmed that you had implemented the changes, you still refer to many authors in the format “Author 1 & Author 2 showed that” (instead of “and”) which is not in line with APA7 guidelines (& is only used when citing in parentheses). Please revise these references to ensure compliance with APA7 formatting. I am sending a separate email with a citation report that will list other issues that need to be addressed.

SUMMARY

In sum, we appreciated your efforts to solidify your results through further analyses and especially through a new study, which is a major strength. At the same time, we were not satisfied with the depth to which you engaged with our previous comments. Therefore, I invite a further revision of the paper, which I would like to receive within a month. Please note that there will not be room for further major revisions in the next round, and I therefore urge you to carefully respond to our comments and implement the necessary changes. I am confident that you can do so and still see clear potential in this research. If you should require additional time or clarifications, please do not hesitate to reach out to me.

Best,

Jonas R. Kunst

Reviewer 1:

The authors achieved many positive changes in the manuscript. First and foremost, it is absolutely applaudable that the authors gathered new data, completing a second study with full randomization between participants, new scales and a more reliable CTAC scale. I also appreciate their response and inclusion of the discussions surrounding different forms of inoculation (e.g. discussing differences between specific inoculations and broad-spectrum inoculations). I also appreciate that the authors added a signal detection approach, which nicely complements their methods.

There are some instances where the authors did not really respond to the comments of the first review, yet at the same time also did not provide sufficient arguments why they did not do so:

- They still refer to priming interventions as useful interventions, saying that they "have been shown to reduce scores on unfounded conspiracy beliefs questionnaires, with small-to-medium effects (d = 0.30-0.56)" (p. 2). They do refer to failed replications later on in the text (in one sentence compared to three sentences about the original studies), but the fact that they still pick studies with small sample sizes to report medium size effects while failing to take into account that large-scale-(i.e., large and representative sample sizes) did not find this effect (and not even balancing the reported medium effects with the effects found by the large samples) portrays to me that the authors fail to acknowledge the underlying problem: That the original studies most likely are built on artifacts / false true findings (made possible also through small samples). The best way to address it would be to either not report the studies that failed to replicate, or to report it in a way showing that while previous studies might have conveyed that they are successful, newer approaches show that they are not (the critique of difficult-to-read fronts might also already brought up here).

- the reliability details for the two subscales of the CTAC are still missing (i.e. they should be included on p. 13). For Study 1b, the reliability of CTAC as well as the reliability for the two subscales is still missing.

- the overall reliability of the CTAC is still low and it does not seem that the authors tried to improve it (e.g. through taking out some of the items). As it stands, the results of Study 1 are still hard to interpret.

- at some instances, effect sizes are still missing, e.g. when reporting the power analyses, see e.g. p.7 "1095 were required for [delete!] to detect an effect size of 0.1 with 80% power" -> is that d or f-squared?

Then there are some instances where the authors added some text that could be improved.

- I appreciate that more details (also theoretical details) about the CTAC are added on page 6. But now the authors claim that they are able to “measure the underlying cognitive process of conspiracy thinking”, and measuring “why” a person believes in a conspiracy theory. While I could see this as a potential goal of the authors, it seems to be a vast overstatement claiming that this is done with the CTAC. First of all, there are many reasons “why” a person believes a conspiracy theory (e.g. personal benefits, social circles) that might be unaffected by discernment. Furthermore, while the authors may show that individuals trained in discernment receive lower scores on the conspiracy scales, this does not necessarily mean that (failed?) discernment is an underling cognitive process of conspiracy thinking.

- “the average number of failed attention checks” portrayed as e.g. M = 0.01, SD = 0.09 are really hard to interpret…. It would be easier to mention absolute numbers (i.e. N of people who failed one or two attention checks).

- The reporting of the simple comparisons of Study 2 for the different conspiracy scales is important, but somewhat hard to read… I was wondering whether there is a way to also portray this information, e.g. in a graph outlining which conditions are different and which ones are not.

- I am not sure the JICT can be easily counted towards a general (“trait-like” measure of conspiracy beliefs). After all, it seems to measure exactly what the authors aim to do (i.e. discernment between potentially true and false conspiracy theories).

- I overall think that the tone of the language in the Discussion could be somewhat more humble, also discussing limitations. For example “[we] found that condition [the Discernment condition] to be the most effective intervention”. -> this statement fails to acknowledge that even the Discernment condition only found an effect on some measures, and that in many cases other interventions produced similar results, though only on subscales. Plus, the low reliability should also warrant more caution.

- I was trying to compare results of Study 2 to the preregistration, but, unfortunately, the authors did not mention one. This seems a bit odd to me as the predictions seem to have been somewhat clear

- I replicated the findings with the data and provided code, but I got different results for the ANOVA in the plausible CTAC subscale, Study 1a (p. 18)

Minor things:

- in the Abstract it should be clear that by “Encouraging discernment…” the authors refer to the condition mentioned earlier

- there still are instances where the language could be improved. E.g. p. 2, line 14. “While misinformation and conspiracy beliefs are distinct (…), they share commonalities in their goal of challenging unfounded beliefs” - > is this the goal of

the beliefs or the researchers studying it? This sentence is unclear. The issue of language applies throughout the manuscript.

Overall, I still think there is value in the manuscript. I like the general idea and the second study make the manuscript far more convincing, but there are still some important questions (some of them still remaining from the first time).

Reviewer 2:

I enjoyed reading the revised version of the manuscript titled “Discernment or Blind Scepticism: Comparing the Effectiveness of Four Conspiracy Belief Interventions”. I was pleased to see that the authors engaged in substantial efforts to address the points raised in the reviews, which I believe greatly improved the quality of the manuscript. Particularly the addition of discernment analyses as well as the additional Study 2 substantially improved the manuscript. However, several issues remain that I recommend being addressed before suggesting the publication of this manuscript. Below I summarise several points I deem important to address and several minor suggestions and remarks. Some of these points pertain to my previous comments, and others to issues addressed by the other reviewer and the editor.

Below is a summary of my main concerns.

Methods:

1. Discernment analyses: I believe the addition of SDT analyses is a great addition to the paper. However, I have additional remarks and suggestions regarding these analyses:
    1. It would be insightful to conduct SDT analyses on the JICT outcome measure in Study 2.
    2. I do not find the argument the authors present for why the factual items added to the IEUB in Study 1 are not suitable for discernment analyses fully convincing (page 15, lines 21-24). It is argued that these items were used to account for response bias, but no analyses testing response bias are reported. Also, while I understand the authors’ point that these very general factual statements are very different from the conspiratorial ones, they should still allow to test whether the tested interventions induce response bias in the form of scepticism to *any* information. I believe that Signal Detection Theory offers a measure of response bias that can be computed independently of *d’*, namely *c* (see Batailler et al., 2022; *Perspectives on Psychological Science*). While I am no expert when it comes to SDT analyses, reporting SDT analyses using *c* as a measure of response bias for the aforementioned measure, but also for other measures allowing for SDT analyses, could provide additional interesting insights.
2. Reliability of scales: The issue of the low reliability of the CTAC in Study 1 that was raised by the editor and the other reviewer was not addressed fully. While

the authors added Study 2, which replicates the results of Study 1 with a more reliable version of the scale (even though the reliability of the "plausible" subscale does also not have very good reliability in this study), this does not change the reliability issues with the scale in Study 1. I encourage the authors to follow the detailed steps outlined previously by the editor to tackle this issue in Study 1.

3. I appreciate that the authors made the syntax for their main analyses available online (although annotating the syntax would be helpful for readers). The outcome variables that seem to have been used for these analyses appear to be ones where outliers have been excluded, based on the variable names and missing values in these columns in the datasets. If outliers were removed, I urge the authors to report this transparently in the main manuscript (merely mentioning this in the preregistration is not sufficient).

Reporting:

1. Currently, the average number of failed attention checks per condition is reported for all studies, which I find somewhat unusual and difficult to interpret. I would find it easier to interpret the total number of participants that were removed due to failed attention checks per condition. Additionally, if these exclusions differ substantially between conditions, this should be mentioned as a limitation.

2. The reporting of attention check exclusion criteria for Study 1 is somewhat unclear to me (page 17, lines 1-6: "*Inoculation and Active conditions, comprehension attention checks were used to ensure that participants had read the infographics properly (e.g. asking participants to describe what the "who benefits fallacy" was, immediately after they had seen the related infographic). Participants who failed at least two of these attention checks were excluded from the analysis. Additionally, participants were excluded if they failed at least one of the Inoculation and Active attentions checks, suggesting that they did not properly read the infographic material.*"). It is not clear right away which attention checks the last sentence refers to, if not the previously mentioned ones, which should be clarified. Additionally, in the lines after that, the authors mention attention checks for the other conditions without mentioning whether the same exclusion criteria applied, which I believe is also not specified in the preregistration. The authors should clearly mention which exclusion criteria were applied for which conditions, in line with, or deviating from, the preregistration, and if these criteria differed between conditions, justify the decision. This recommendation also applies to Study 2.

3. In reporting Study 1b, the authors mention that they controlled for age and gender in their analyses, but no such analyses are reported in the manuscript or supplementary material. I encourage the authors to clearly specify which models including these covariates were fit to the data, provide the code for these analyses in the online supplement, and, optionally, report the results in the supplementary material.

4. When reporting results for the IEUB on page 25, it is not mentioned what kind of model was fit to the data. I assume the same model was used as for the other outcome variables, but this should be specified.

Clarity:

1. I appreciate that the authors clarified some sentences and paragraphs that were unclear previously and corrected some errors. However, I believe the manuscript still requires some proofreading. Hereafter is a non-exhaustive list of some errors and style issues I spotted:
    1. Page 1: The consequences of belief in (unreasonable) conspiracy theories are summarised twice in two separate paragraphs, which is somewhat repetitive and could be condensed.
    2. Page 22, line 20 should read “affected” instead of “effected”.
    3. Page 6, line 23: A full stop is missing.
    4. The (rather frequent) usage of “as such” throughout the paper sounds a bit odd in some sentences.
    5. Statistical notation symbols such as d’ or z should be italicised.

Hereafter is a summary of additional minor issues and comments.

Theoretical background:

1. Presentation of prior research on priming interventions: I believe the presentation of previous research on priming interventions as “promising” could be further toned down. While I appreciate that the authors caveated the presentation of prior research by briefly mentioning failed replications/replications showing much smaller effect sizes, I believe the fact that more recent (direct) replication studies with higher power than the original ones failed to replicate these effects warrants making these inconsistencies more salient/presenting these studies in a more critical way. More specifically, I would not present them as “promising” interventions, as is currently the case, but instead highlight these inconsistencies, which I believe would be an argument for testing these interventions in the presented studies that is equally strong as the current presentation of these interventions as “promising”. The literature on such priming interventions is already covered more critically in the discussion section, which I appreciate, but I think this is not reflected in the introduction (which actually also results in a discrepancy between the introduction and the conclusion in the evaluation of this literature).
2. In response to my previous recommendation to discuss why some of the tested interventions appear to induce undiscerning scepticism, the authors added a brief discussion on page 37 that I find to be a valuable addition to the paper. However, the point I raised previously was not directly addressed. I still

believe that a brief discussion of these points would be valuable (see point two in the section on minor issues and comments regarding the theoretical background in my previous review). However, I would also understand if the authors do not feel that their studies allow for an evidence-based discussion of these points and leave the decision to include such a discussion to the authors.

3. With the additional measures of conspiracy beliefs introduced in Study 2, the manuscript now includes quite a few different measures of related, but (partly) distinct concepts. It might be good to explain the relations between these constructs and their differences somewhere. This could be done in the theoretical background, when introducing the measures, or in the discussion of interventions differentially affecting different outcome measures.

4. I find it good that the authors added that misinformation and conspiracy research (particularly on interventions) are related. However, the sentence "*While misinformation and conspiracy beliefs are distinct (De Coninck et al., 2021), they share commonalities in their goal of challenging unfounded beliefs.*" on lines 12-14 of page 2 could be misinterpreted. I would clarify that it is *interventions* that are similar in their goal of challenging unfounded beliefs or, alternatively, that both beliefs in misinformation and conspiracy beliefs are instances of unfounded beliefs. However, labelling conspiracy theories as unfounded contradicts the concept of plausible conspiracy theories. It could be argued though, that they may be formed in a similar way (as they share similar antecedents, correlate strongly, etc.), which would avoid this contradiction (and this would also be in line with your revised conceptualisation of the construct the CTAC taps into).

Methods:

1. When reporting power analyses, it should be reported which effect size measure was used. While in response to a comment by Reviewer 1 it was noted that effect sizes were calculated in terms of *F*, 1) this is not reported in the manuscript and 2) I believe the authors refer to the effect size *f* (lowercase), which is not to be confused with an *F* (uppercase) statistic.

2. It would be good to add the response to my remark regarding the unclear reporting of how the CTAC is measured (i.e., that 1 point per correct answer can be obtained) to the text of the manuscript (I could not find it there). I find the current description ("Scores ranged from a possible one to eight (min=1, max=8)" page 13, line 5) to be not very precise.

Reporting:

1. Reporting correlations between the (outcome) measures reported in each study would be interesting for readers. Doing so in the supplementary material should suffice.

2. The SDT analyses were exploratory in nature, which should briefly be noted.

3. I would check if the p-value on page 19, line 1 actually equals exactly .00. If not, it should be reported as < .001.
4. When reporting results for the total CTAC scores (Study 1b: page 23, first paragraph; Study 2: page 30, line 22), it should be explicitly mentioned that this refers to the total CTAC scores.

Discussion of results:

1. I would find (very) short interim discussions of Study 1 and Study 2 results after the results are reported helpful.
2. In Study 2, there were no effects of the discernment intervention on JICT (which is more of a belief measure), but there were effects on CTAC (which relates more to the way of thinking that may result in beliefs). This suggests to me that CTAC indeed taps into a different construct than beliefs and I find the fact that this is captured in the presented results highly interesting. It seems that discernment interventions may be particularly relevant for tackling thinking processes (which entails the process by which some information is discarded and some accepted; i.e., discernment) rather than beliefs once they are formed (which are a result of these thinking processes). I leave it up to the authors whether to implement this suggestion or not, but I find this a very interesting finding that could be discussed in more detail.

I enjoyed reading the revised version of this manuscript and believe that it will be a highly interesting and impactful paper once the most important issues I mentioned have been tackled. I wish the authors all the best for the next steps with this manuscript.

Round 3

Dear Authors,

Thank you for your diligent attention to our previous comments. The paper has improved significantly. Given the substantial changes, I have sent the manuscript to the reviewers for another and hopefully last round of evaluation to ensure it reaches its full potential.

The reviewers have highlighted several points that require additional attention. I am grateful to them for their persistent efforts to provide high-quality feedback. In terms of Reviewer 1, please use the “updated” review attached directly after this message because the other version (attached later) was based on an older version of the manuscript. While we are close to finalizing the manuscript, please address the following issues along with the reviewers' comments:

1. When referring to statistically non-significant effects in the abstract and main text, describe them accordingly. For example, instead of "did not reduce general conspiracy ideation" (which requires a direct test of the null hypothesis), write "did not statistically significantly reduce general conspiracy ideation." Please apply this change throughout the manuscript.

2. In the abstract, replace "in this study" with "in this research" to accurately reflect the two studies conducted.

3. Change all instances of "participants that" to "participants who" in accordance with APA guidelines.

4. If possible, provide Figure 1 as an SVG file. Alternatively, supply it in higher resolution.

5. Include figures for the results of Studies 1a and 1b in the main text, as well as a graph for the c scores.

6. In Figure 2, panels B and D seem identical. Please review and correct.

7. Verify and correct the degrees of freedom reported for the ANCOVA F tests. They seem too low.

8. Address the following stylistic issues:

a. Add the missing page number for the quote beginning with: Unlike standard informational inoculations, active inoculations are “a two-way process through which participants engage interactively” (Kiili et al., 2024).

b. There are commas missing in this sentence “Topics in the inoculation related to (i) false experts (Banas & Miller, 2013; Cook et al., 2017), (ii) overcomplicating explanations and making unnecessary assumptions (Banas & Miller, 2013; Stall & Petrocelli, 2023) (iii) identifying logical leaps (Pytlik et al., 2020; Roozenbeek et al., 2022) (iv) the “who benefits” fallacy (Wagner-Egger, 2022) and (v) looking for corroborating evidence (Stall & Petrocelli, 2023).”

c. There are still statistical letters that are not italicized. E.g., p. 14, l.7 (*SD* in italics). Same applies many other places (e.g., p. 21; also note that η^2 is not correctly formatted there, same on p. 25 and 27; see also p.22, p. 31 for formatting issues). **Please check everything thoroughly.**

d. Standardize the placement of full stops in relation to quotation marks. Either within quotation marks or after.

e. Add the missing word in the sentence "…results 3 of the multiple comparison analysis broadly the same as the main ANOVA analysis."

f. Sometimes eta-squared is reported as partial and other times this information is missing. Is this on purpose because it refers to different etas?

g. Ensure consistent spacing around "<" and "=" symbols.

h. Replace the comma with a decimal point in the correlation value in this sentence: "The Implausible and Plausible subscales were found to negatively correlate with one another (*r* = -.45, p <,001)."

i. Clarify the meaning of "< .01" in Table 1 (probably a typo) and add the missing 0 before ".03".

j. Remove bold formatting from punctuation at the end of p. 26 and beginning of p. 27.

8. This sentence ends with a comma instead of a full stop: "The CTAC was once again used to measure critical appraisal of novel conspiracy theories. The internal reliability was low for the overall assessment (α = .55) and for the Plausible (α = .50) subscale. The Implausible subscale had acceptable reliability (α = .63),"

l. Correct the punctuation in the sentence on p. 23 ending with "Is this suspicious.?"

m. Remove the period after "As seen in Figure 3".

n. Complete this incomplete sentence regarding the Inoculation condition results: "Those in the Inoculation condition were significantly less inclined to rate plausible conspiracy theories as being likely than those in the Priming (p = .049, d = -0.37) also marginally less likely than Control conditions (p = .053, d = -0.38)."

I **strongly** recommend having another person proofread the entire manuscript to catch any remaining stylistic issues so we don't go in further loops. This step is crucial to ensure we can move forward efficiently with the publication process.

Thank you for addressing the substantive issues in your previous revision. We are now close to reaching the level required for publication in the special issue. Please submit a revised version within one month alongside a revision letter. I hope to be able to make a final decision in the next round.

Best regards,

Jonas R. Kunst

Reviewer 1:

Overall, the authors did a great job in responding to the last review. I especially appreciate the author's discussion of the (null-)findings regarding the priming interventions and like their change in phrasing from "promising interventions" to "popular interventions". Also, the overview of the present Study was done very nicely.

I don't agree with the authors presentation of the JICT as measuring general conspiratorial tendencies, in itself and as a solution or explanation why no effects on that measure were found. First of all, there is a lack of clarity: In the conspiracy literature, the "propensity to think in a conspiratorial manner" typically is referred to as conspiracy mentality, which measures abstract concepts (of which some are also included in the GCBS, so I don't necessarily oppose framing the GCBS as such). However, the JICT measures specific plausible and implausible conspiracy theories. Given that these are specific theories, participants should be more easily swayed on this measure than on measures of conspiracy mentality. This might sound very picky, but recently, there have been repeated calls in the literature to be more precise regarding conspiracy belief concepts, and I think that the paragraph could improve in this regard. Additionally, as mentioned in my first review, I still think that the Discernment intervention and CTAC are relatively similar in terms of reasoning

and answer format. I would love to see this discussed, both as a general limitation, as well as one explanation why the discernment condition was effective on the CTAC but not the JICT. Overall, if the discernment condition indeed improves discernment between plausible and implausible conspiracy theories, I would expect a difference on the JICT, and the fact that it is not found (but only regarding a measure very similar to the intervention), still poses a severe limitation for me regarding this intervention and, for me, questions it superiority regarding the inoculation interventions (which, unlike the discernment condition, at least report lower agreement to the GCBS and the IEUB compared to the priming condition in Study 2; and all of them report lower scores on the IEUB in study 1b). Thus, while I don't want to play down discernment as an intervention potentially being similarly effective to inoculation (while potentially improving critical assessment skills, which still needs to be tested with a DV less similar to the intervention), I would also advise to tone down regarding this superiority claim. I know that there is the impression that authors need to be super-confident and advertising of their own interventions and scales. But overall, I think that a discussion honestly admitting open questions on the long run does more service to the field.

I also understand and agree with removing the analyses for the scales with low reliability to the supplement (which I was not able to access and thus, not able to review, just as a note). The reason for this change (also noted by the authors) is the fact that the results should only be interpreted tentatively. I wish that this tentativeness would also be reflected more in the way the authors talk about these findings. E.g saying that in Study 1b "*we found some evidence* that encouraging discernment improved the ability to critically appraise conspiracy theories" is an overstatement. Discussion of those findings should also be more tentatively for Study 1a and in the Discussion.

Apart from that, I only have minor comments:

1) Regarding the new abstract: Given the similarity to JICT (also assessing "appraisal of both plausible and implausible conspiracy theories") on which the authors found no effect, it would be good to be more precise regarding the CTAC, e.g. classifying it as "judgment regarding novel conspiracy theories"

2) In the abstract, the authors also mention that the "inoculation-based interventions were moderately successful at reducing epistemically unwarranted beliefs but did not reduce general conspiracy ideation". This leaves the reader with the question of what happened to the other conditions, so it would be helpful to include "inoculation-based interventions [but not Priming and Discernment] were moderately successful…"

3) In the letter, the authors mentioned that they calculated the power analysis for the effect size f-squared. However, in the paper on p. 8, they mention the effect size "f" (which is a difference to f-squared). Please clarify.

4) I think the description of the CTAC on page 8 as having the "*ability* to measure the underlying cognitive process of conspiracy thinking" is very bold (while this might be the goal, the current studies do not test whether it is actually doing that)

5) The self-reported enjoyment on p. 17 is missing the SD.

6) I would advise another language check, specifically for those newly added paragraphs, as they included some minor language mistakes

Reviewer 2:

I thank the authors for their efforts to address the issues raised in the reviews, which have substantially improved the manuscript titled "Discernment or Blind Scepticism: Comparing the Effectiveness of Four Conspiracy Belief Interventions". While I do not see many major issues with the current version of the manuscript, I believe there is a host of minor points that should be revised before the publication of this manuscript. Some of these points pertain to issues that have been raised previously, while others are new.

Length of the manuscript: Currently, the manuscript is relatively long and could be shortened. For instance, this could be done by moving some of the verbatim details about attention and comprehension checks to the supplementary material or shortening it. Furthermore, the presentation and analysis of enjoyment measures could be moved to the supplementary material, as they are not crucial to the main line of argumentation, I believe (also note that the point in the discussion about how more engaging interventions do not run the risk of decreasing the willingness to participate because there were no differences in measured enjoyment is somewhat of a stretch: enjoyment is measured after participation and the willingness to participate precedes participation). Another aspect that could be shortened is the interpretation of results from analyses involving variables with reliabilities that are too low to allow for a meaningful interpretation (which is somewhat of a contradiction in the current version of the manuscript). Regardless of whether the authors decide to shorten these specific aspects, I think it would be important to shorten the manuscript somehow to make it an easier read.

Attention checks:

- It makes sense that there are different percentages of failed attention checks per condition if different (numbers of) attention checks were used in different conditions – therefore, I find the analysis comparing the total or mean number of failed attention checks per condition not very informative. It would make more sense to me to just compare the number of failed attention checks across conditions for the attention check items that are the same across all conditions (if necessary, for individual items separately). If there are substantial differences between conditions in failure rates for these items, this suggests that there might be a different underlying issue than the one discussed by the authors.

- The authors state that the results including observations with failed attention checks are broadly the same, but there are actually relevant differences: e.g., for Study 1a, the difference between intervention groups on IEUB scores is not significant when including failed attention checks, but it is significant when excluding them. To tackle this without adding to the length of the manuscript, the authors could simply state that there are some differences which are reported (and elaborated on) in the supplementary material.

Formatting:

- Some of the CTAC figures look like they still include values below 0 (which should not be within the range of possible values unless I am mistaken), which should be corrected.
- I think there might be some issues with the formatting of headings, as some of the headlines that should clearly be at different levels (e.g., the heading for Study 1a and the heading for the methods section for Study 1a) are formatted in the same way.

Presentation of discernment analyses and measures:

- I do not fully understand (the relevance of) the argument the authors make in the discussion section about the JICT not being suitable as a measure of discernment. To me that is not (only) an empirical question as the authors appear to imply based on their argument that is solely based on correlations with other measures, but mainly a question about the nature of the measure. As long as it measures two discernibly different types of stimuli, it should be suitable as a discernment measure, regardless of correlations with other measures or between subscales (unless they are extremely high). To me this seems to be a minor technicality, so unless the authors have good reasons to include this point that I might have overlooked, I think it could be omitted.
- I appreciate that the authors added analyses for the SDT measure of response bias (i.e., c). The interpretation of the results from these analyses could be connected more strongly to one of the authors' goals, namely to "establish whether interventions are teaching participants how to identify unfounded conspiracy theories, rather than simply a response bias to dismiss all conspiracy theories".

Summary of findings:

- In the abstract, the authors summarise the focal effects of different interventions, but they summarise effects of different interventions on different outcomes – instead of this (or in addition) it may be interesting to summarise how the different interventions perform in terms of the same outcome variables if the word limit allows for that.
- The manuscript presents many different measures, analyses, and results, which is good. However, this made it difficult for me to follow the main line of argumentation and see what the main take-aways are. This probably also relates to the length of the manuscript and some style issues, so tackling

these points could help making the manuscript more engaging and easier to follow.

Statements that could be toned down or should be more specific:

- On page 1, priming interventions are still portrayed as “promising”. I understand that this appears to be the conclusion of the O’Mahony et al. review, but in line with previous comments, these interventions can and should not be portrayed as promising given the current state of evidence.
- In the discussion on page 39 (line 14), instead of stating that the current research provides evidence that some interventions have “no effect” (which would require an equivalence test providing evidence that there is indeed no effect), stating “no statistically significant effect” would be more appropriate.
- Some statements in the discussion could be toned down, such as statements about the discernment intervention being the most promising. While this appears to be the case for some measures, other conditions perform better on other outcome measures.
- This relates to a point also raised by the editor about how positive and negative effects of established interventions were presented as “approximately equal”. If the authors refer to effect size, then they should provide empirical evidence for these effects/coefficients being of a similar size (e.g., a linear hypothesis test comparing a model whereby both coefficients are constrained to be of the same size to an unconstrained model). Otherwise, these claims should be omitted or toned down.

Preregistration:

- On page 31 (lines 15 and 16), the authors state that they exclude participants based on their preregistration, despite the fact that they state earlier that Study 2 was not preregistered.
- On page 33 (line 1), the authors state that their SDT analyses are exploratory. However, since Study 2 was not preregistered, all analyses should be considered exploratory, so either none or all analyses should be labelled as such, not just some, which could be interpreted to suggest that other analyses were confirmatory.

Errors and style issues (note that this is not a complete list):

- Some phrases, while fully understandable, sound a bit odd:
    - The second sentence in the abstract reads “little effect on increasing critical appraisal”. Something like “little effect on critical appraisal” or “did not increase critical appraisal” would sound better.
    - Page 26, lines 13 and 14: “in comparison to the Priming group, who has a bias”. This could be reframed as “which shows a bias”.
    - Page 3, line 2: “robust samples” sounds odd.

    - Page 3, lines 6 and 7: “have failed to meet manipulation checks” could be rephrased as “have failed to manipulate the intended construct” or something similar.

- Abstract, line 8: the authors state that the discernment condition encouraged “reasoning”, which is very broad and could also apply to the other conditions in a way, so this could be rephrased to be more specific.
- Page 2, line 19: the two sets of parentheses could be combined into one.
- Page 6, line 16 should read “a lack *thereof*”.
- Throughout the manuscript, the capitalisation of “van der Linden” is inconsistent.
- Spacing before and after = is inconsistent throughout the manuscript.
- Page 21, line 15: “more *involved* interventions” is incorrect.
- Page 21, line 18: “judgments of critical thinking skills” is not what was measured; “judgments of” can be omitted.
- Page 28: there are some instances of the letter *a* being used instead of alpha.

I enjoyed reading this version of the manuscript and believe that it has the potential to be a very good paper, given the aforementioned issues are dealt with. I wish the authors all the best for the next steps with this manuscript.